Using Docker to configure a cluster of Hadoop and Spark nodes to store and process health data within the African American community

Dylan Bent
Department Electrical Engineering
Florida International University
United States
dbent011@fiu.edu

***Abstract*—The purpose of this project was to conduct a study on the health outcomes of African American communities and possible adverse factors using distributed computing.**

I. Introduction

The African American population has witness high mortality rates as compared to their white counterparts. Although trends showed a decline in mortality rates in the early 2000s, 2020 and the introduction of COVID-19 saw a rise in mortality rates within the African American community [1]. In addition to the rise of COVID-19, the African American community has reported high levels of obesity, with four out of five African American women being obese [2]. Obesity increases the likelihood of cardiovascular illness [2]. Other cardiovascular illnesses common to the African American community also include asthma. Asthma is more common in the African American community than any other group in the United States [3]. African Americans are up to forty percent more likely to suffer asthma than White Americans. These cardiovascular conditions contribute to higher mortality rates and exacerbate the symptoms caused by COVID-19. Other illnesses, like diabetes also complicated the mortality rates in the African American community. The rate of diabetes in the African American community is only second to that of the rate in Indigenous American communities [4]. Considering that diabetes is the eighth leading cause of death in the United States, it's reasonable to assume this has a significant impact on the mortality rate in the African American community. This mortality rate is also impacted by lack of resources including access to affordable healthcare and insurance [5]. Financial burdens make it difficult more many Americans to cover the cost of healthcare, and this also negatively contributes to adverse health outcomes in the African American community.

II. Dataset

The dataset used in this research comes from the Centers for Disease Control and Prevention (CDC) along with the U.S. Department of Health and Human Services (HHS). Together these government agencies created the Minority Health Social Vulnerability Index (SVI). This data helps to determine geographically where Americans are medically vulnerable and possible factors and outcomes. The dataset used in this report is from 2020 and includes various health outcomes such as obesity, diabetes, respiratory diseases, and deaths as the result of cardiovascular illness. Factors of adverse health outcomes includes those such as housing insecurity, low educational attainment, access to medical institutions among others.

III. Design

In order to efficiently store, clean, and process the data in the SVI dataset, it was determined that distributed computing could greatly increase the efficiency of the process. To store the dataset, the CSV file was uploaded to a Hadoop Distributed File System, where the file was broken into smaller blocks and replicated amongst three nodes. This ensures that the data is highly consistent and partition tolerant. The distributed dataset is then imported into a Spark cluster of three worker nodes for processing. PySpark is used to transform the dataset into a DataFrame data format. To connect to this Spark cluster, a Jupyter notebook is used to run the computational commands and handle output. To orchestrate the entire process, Docker is used. Hadoop and Spark nodes are containerized, configured, and automatically deployed. Using a Spark cluster helps to partition the data and run computational processes on smaller chunks of the data in parallel. This process is automated and handled by PySparks DataFrame interface. Once the results are created they can be outputted within the Jupyter notebook as well as saved to the Hadoop Distributed File System where they can be retrieved.

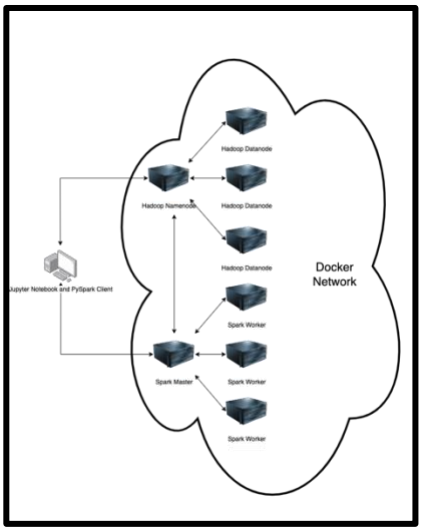

Figure 1.

Figure 2.

IV. RESULTS

Several data visualizations were prepared based on the data. One is a correlation heatmap to better understand the statistical correlation between health factors and outcomes (Figure 3). Notably the data shows a distinctive relationship between poverty and poor health outcomes in the African American community. In addition, it reveals that the lack of hospitals and pharmacies leads to greater risk of respiratory diseases and cardiovascular-related deaths. Lack of internet which predicts poor cardiovascular health also shows to improve rates of proper fat percentage as well as lower rates of diabetes. The heatmap also reveals that housing crowdedness has less of an impact on respiratory health than anticipated.

The second set of data visualizations map the population density of African-Americans per state as well as the rate of health issues in each state including obesity, diabetes, respiratory disease, and cardiovascular-related deaths. After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

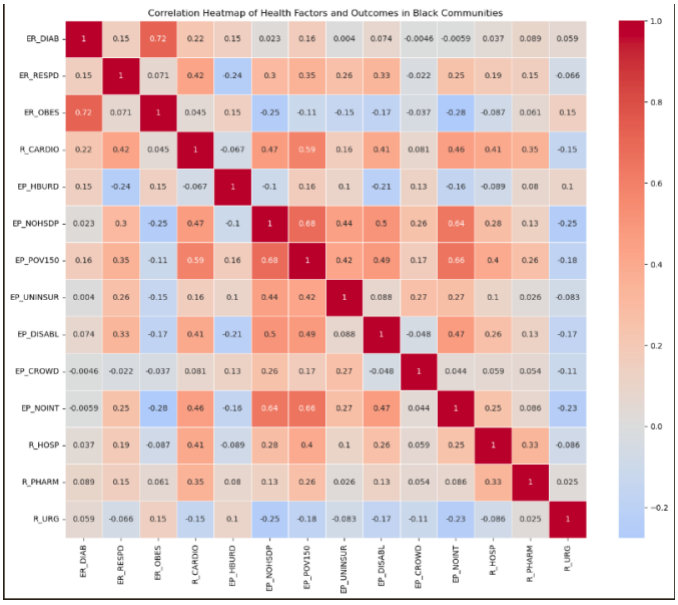

Figure 3.

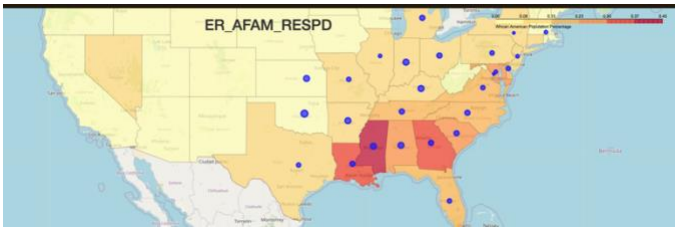

Figure 4.

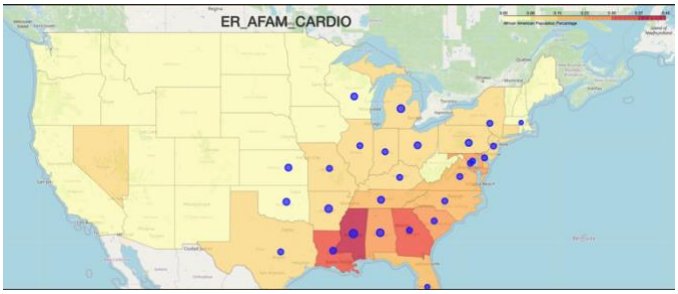

Figure 5.

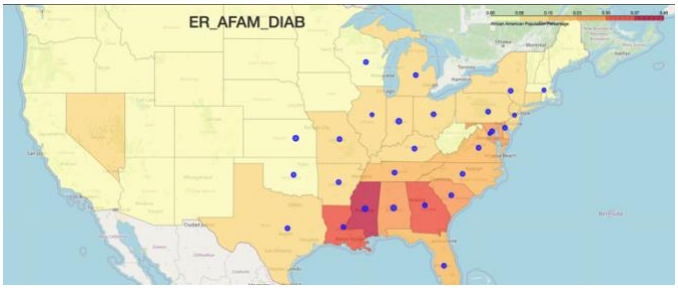

Figure 6.

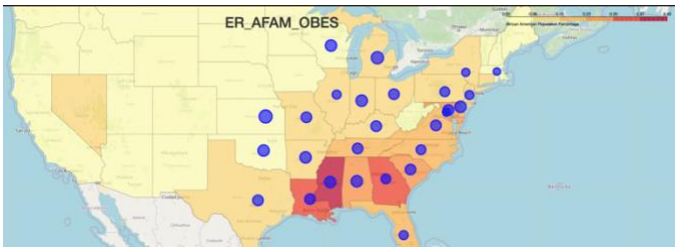

Figure 7.

The figures 4 through 7 show high rates of poor health outcomes in the southeast of the United States. These regions, coincidentally align with higher rates of African American populations. This aligns with the findings that African Americans suffer from higher rates of obesity, diabetes, respiratory disease, and cardiovascular-related deaths.

V. Summary

The data visualizations in this paper were able to be processed efficiently used distributed computing systems. The ability to interact with computer systems from a simple Jupyter Lab terminal aided in making corrections quickly without th need for heavy infrastructure re-development. Despite the significance of the findings in the dataset, there needs to be more specific datasets oriented at specific census-defined groups. The statistics for African Americans in this report were determined by factoring the size of African American communities per county, and the rate of negative health outcomes in those same counties. Isolating the datasets by ethnic groups will make it easier to create more accurate statistics.

References

https://github.com/bentd/bigdata-research-project-2024

[1] César Caraballo, M. (2023, May 16). *Mortality and years of potential life lost in the US black population*. JAMA. https://jamanetwork.com/journals/jama/article-abstract/2804822

[2] *Obesity and African Americans*. Office of Minority Health. (n.d.). https://minorityhealth.hhs.gov/obesity-and-african-americans

[3] Asthma in the Black Community. (n.d.). https://www.nhlbi.nih.gov/sites/default/files/publications/asthma_in_black_community_fact_sheet.pdf

[4] *Statistics about diabetes*. Statistics About Diabetes | ADA. (n.d.). https://diabetes.org/about-diabetes/statistics/about-diabetes

[5] Samantha Artiga, L. H. (2024, February 22). *How present-day health disparities for black people are linked to past policies and events*. KFF. https://www.kff.org/racial-equity-and-health-policy/issue-brief/how-present-day-health-disparities-for-black-people-are-linked-to-past-policies-and-events/